# Detecting adaptive evolution based on association with ecological gradients: Orientation matters!

Eric Frichot[1], Sean D. Schoville[2], Pierre de Villemereuil[3], Oscar E. Gaggiotti[3,4], Olivier François[1]

[1] Université Joseph Fourier Grenoble, Centre National de la Recherche Scientifique, TIMC-IMAG UMR 5525, 38042 Grenoble, France.

[2] University of Wisconsin-Madison, Department of Entomology, Madison, WI, U.S.A.

[3] Université Joseph Fourier Grenoble, Centre National de la Recherche Scientifique, Laboratoire d'Ecologie Alpine UMR 5553, 38042 Grenoble, France.

[4] University of St Andrews, Scottish Oceans Institute, East Sands, St Andrews, Fife, KY16 8LB, UK.

## Abstract

Population genetic signatures of local adaptation are frequently investigated by identifying loci with allele frequencies that exhibit high correlation with ecological variables. One difficulty with this approach is that ecological associations might be confounded by geographic variation at selectively neutral loci. Here we consider populations that underwent spatial expansion from their original range, and for which geographical variation of adaptive allele frequency coincides with habitat gradients. Using range expansion simulations, we asked whether our ability to detect genomic regions involved in adaptation could be impacted by the orientation of the ecological gradients. For three ecological association methods tested, we found, counter-intuitively, fewer false positive associations when ecological gradients aligned along the main axis of expansion than when they aligned along any other direction. This result has important consequences for the analysis of genomic data under non-equilibrium population genetic models. Alignment of gradients with expansion axes is likely to be common in scenarios in which expanding species track their ecological niche during climate change while adapting to changing environments at their rear edge.



**Keywords:** Ecological association methods, Local adaptation, Range expansions, Ecological gradients, Genome scans.

# Detecting adaptive genetic responses using ecological gradients

The fossil record is replete with examples of species modifying their geographical distributions following environmental change (Blois & Hadly 2009, Comes & Kadereit 1998). Based on evidence from the past, range shifts are commonly viewed as an expected response of species to climate change (Parmesan & Yohe 2003). In addition to range modifications, changing conditions and natural selection can also trigger genetic modifications allowing species to adapt to new local environments encountered during migration (Davis & Shaw 2001, Davis et al. 2005, Jump & Penuelas 2005). Under these conditions, researchers have suggested that gene frequencies may change gradually as natural selection acts on standing genetic variation or new mutations to favor adaptive phenotypes (Hermisson & Pennings 2005, Hancock et al. 2010, Pritchard et al. 2010). In addition, natural selection may cause shifts in allele frequency at multiple genetic loci simultaneously, leaving subtle signatures of adaptive genetic change (Vitti et al. 2013).

Detecting gradual parallel genetic change is a difficult challenge (Pritchard et al. 2010), and range expansion scenarios may complicate the identification of adaptive alleles among selectively neutral polymorphisms. Range expansions can generate extreme genetic drift in the direction of colonization, driving allele frequencies close to fixation in a pattern that mimics the signature of selective sweeps (Edmonds et al. 2004). These results imply that traditional outlier methods based on allele frequency differentiation may be inappropriate to detect genetic signatures of local adaptation in species that underwent range expansion

in the past. Allele frequency differentiation tests indeed lack power to detect soft sweep signatures, and they are prone to high false positive rates in this situation (Teshima et al. 2006, Hermisson 2009).

A way to investigate signatures of local adaptation when selective alleles have weak phenotypic effects is by identifying loci with allele frequencies that exhibit high correlation with ecological variables (Joost et al. 2007, Hancock et al. 2008). Genome scan methods based on association of loci with ecological gradients assume that environmental factors vary throughout geographical space, and provide good proxies for unobserved selective pressures. Ecological association methods have performed well in simulation studies, often detecting adaptive loci when outlier allele frequency differentiation tests have failed, as in cases, for example, where selection varies geographically (De Mita et al. 2013), or when adaptive phenotypes evolve as polygenic traits (de Villemereuil et al. 2014).

An unanswered question is whether ecological association methods are impacted by the shape and orientation of gradients used as proxies for selection. Answering this question is essential to the interpretation of ecological association tests and to the use of ecological predictors that would produce the smallest proportion of false positive associations. In this study we considered a fictive species in which adaptive allele frequency gradients correlate with the geographical variation of some known environmental variables after the species expanded from its original range. For empirical examples, consider the evidence for genetic change associated with the ongoing range expansion of the bank vole (*Myodes glareolus*) in Ireland, or with the recent range expansion of the British butterfly

(*Aricia agestis*) in response to climate change (Buckley et al. 2012, White et al. 2013). Under the situation described above, we argue that the performance of genome scan methods based on associations with ecological gradients will depend on the orientation of the gradient relative to the expansion axis as well as the test used. We provide evidence of spurious association at selectively neutral alleles for some gradients, and show that the most favorable case is when the gradients align along the direction of expansion. These results provide guidelines for researchers to analyze results when testing genetic data with ecological association methods.

## Ecological association methods identifying genomic signatures of local adaptation

Genome scans based on correlations of allele frequencies with ecological gradients use multiple statistical tests to detect significant association at each locus. The list of loci showing associations with local environment are considered as candidate loci potentially targeted by selection. These associations are commonly tested using regression models (Joost et al. 2007). For example, regression models have been employed to identify ecologically relevant loci in humans (Hancock et al. 2008, Fumagalli et al. 2012, Frichot et al. 2013). Jones et al. (2012) used ecological correlation methods in their comparison of marine and freshwater sticklebacks to detect loci associated with habitat. Eckert et al. (2010) used regression methods to identify loci linked to climatic gradients in loblolly pines.

More specifically, statistical methods that evaluate the association of gene frequencies with ecological gradients can be classified into 3 main categories.

Some of these methods include corrections for confounding effects due to population structure, whereas some others do not. The first category of methods tests for correlations using linear or logistic regression models or simple Mantel tests (Joost et al. 2007). These methods are appropriate for continuous populations or populations interconnected by high rates of gene flow. In other contexts, these simple regression models generate large numbers of false positive associations (Schoville et al. 2012, De Mita et al. 2013, Frichot et al. 2013).

A second category of methods explicitly considers geographic structure in the data, and corrects for confounding effects created by shared demographic history and patterns of isolation by distance. Those models estimate the effect of the ecological variable on allele frequencies allowing for statistical dependence in residual terms. Interestingly, correction for confounding effects in ecological association methods relies on principles that are similar to phylogenetic comparative methods (Grafen 1989, Harvey & Pagel 1991). Phylogenetic comparative methods use information in phylogenetic trees to test for correlated evolutionary changes in two traits. In ecological association tests, confounding effects are corrected by introducing a known covariance matrix that models allele frequency dependencies. Evidence for local adaptation at a specific locus is then evaluated by testing a null model with this particular covariance structure. Our analogy with phylogenetic comparative methods postulates that background correlation in ecological association methods could be modeled on the basis of geographic or population genetic distance instead of phylogenetic distance (*cf.* Felsenstein 2002).

For example, Poncet et al. (2010) used a generalized estimation equation

model which assumes a covariance matrix in which nearby individuals are genetically more similar than individuals located farther apart. Another approach estimates the empirical covariance of allele frequencies among populations and uses it as the null model (Coop et al. 2010). This approach was implemented in the computer program `BAYENV`, and uses the full covariance matrix of allele frequencies. The `BAYENV` model can be extended to consider low rank approximations of the covariance matrix. It is then similar to using a regression model in which a fixed number of principal components of the data matrix are included as fixed effects in the model.

A third category of methods has been inspired by genome-wide association studies and mixed models (*e.g.* Yu et al. 2006, Frichot et al. 2013, Yoder et al. 2014). Association between ecological gradients and allele frequencies are tested while estimating the effects of unobserved latent factors. In mixed models, the latent factors include background levels of population structure due to demographic history or background genetic variation, and the fixed effects model the correlation between allele frequencies and the observed selection gradients. The mixed model approach implemented in the software `LFMM` 2.1 (Frichot et al. 2013) proved to be among the most reliable approaches in a recent evaluation of several genome scan methods (de Villemereuil et al. 2014).

In this study we used `LFMM` in conjunction with two other ecological association tests based on regression methods to investigate the confounding effect of range expansions on our ability to identify genomic signatures of selection. The two methods used a linear regression model without correction, and a linear regression model in which a fixed number of principal components of the data

matrix are included as fixed effects. For linear regression methods, we used classical testing procedures based on $z$-scores and $t$-tests. Using `LFMM`, the number of latent factors, $K$, was chosen on the basis of the empirical distribution of locus-specific $P$-values after each program run. More specifically, we ran `LFMM` five times for each value of $K$ with run-lengths of 10,000 cycles and burn-in periods of 5,000 cycles, and we combined the $P$-values and the $z$-scores resulting from each run using the Fisher-Stouffer method (Brown 1975). Following Devlin and Roeder (1999), we obtained a genomic inflation factor (GIF) after computing the median of the squared (combined) $z$-scores for each $K$, divided by the median of the $\chi^2$ distribution with one degree of freedom. We selected the smallest value of $K$ for which the GIF value dropped below 1, using it to correct the test $P$-values.

## Spatial simulation of neutral and adaptive allele frequencies

We considered a fictive species that underwent a range expansion 1,000 generations ago. For this species, we simulated a demographic model in which a rectangular area was colonized from a unique source population located south of the area, and we considered population samples from the whole species range at the end of colonization.

In our simulations, the main axis of expansion was oriented in the northward direction. We used the Haldane cline model to simulate geographic variation at adaptive loci based on ecological gradients (~~Haldane 1948,~~ see below). A reference ecological gradient was defined to be parallel to the main axis of expansion. Then the axis of the reference gradient was rotated by angles of 11.25

degrees from the original position. We considered a total of 17 distinct angles ranging from -90 to +90 degrees. See Figure 1 for a representation of our simulation framework. An angle of 0 degree represented a selection gradient parallel to the main axis of expansion. We simulated independent genetic variation at 4,900 neutral and at 100 adaptive single nucleotide polymorphisms (SNPs). Our simulated data sets contained low percentages of true associations with ecological gradients (2%). We also simulated SNP data using 4,500 neutral and 500 adaptive loci.

Data sets consisting of selectively neutral multi-locus genotypes were created using the computer program `SPLATCHE` (Currat et al. 2004). Range expansion scenarios were implemented using non-equilibrium stepping-stone models based on a regular array of 165 demes organized in a rectangle of size 11-by-15. a rectangular area was colonized from a unique source located south of the area (Figure 1). For each deme, the migration rate was equal to $m = .4$, the expansion rate was equal to $r = .4$, and the carrying capacity was equal to $C = 100$. The "density overflow'' option was used to spread the source population over $\approx$ 8 demes.

Four genotypes were sampled from each of the 165 demes for a total number of 660 genotypes. To create associations between loci and ecological gradients, we linked allele frequencies to ecological gradients by using Haldane's transform (Haldane 1948). The Haldane transform simulates a geographic trend, *i.e.*, continuous variation through geographic space, that reproduces clinal allele frequency patterns as expected under spatially varying selection intensities. In addition, we used a model of correlated residuals that generates the same

background population genetic structure at adaptive loci as observed at neutral loci. To implement it, we introduced residual errors based on the empirical covariance matrix of the neutral loci (Coop et al. 2010). The shape parameter for Haldane's clines was set to mimic weak selection, not easily detectable using classical population differentiation methods. To check this, we computed the first axis of a principal component analysis for a typical set of neutral SNPs (Figure S1A). This axis clearly separated populations defined at the right and left of the expansion axis. For all data sets, we computed the empirical distributions of $F_{ST}$ for populations defined at the right and left of the expansion axis. Running tests with statistical power greater than 80%, we found that the false discovery rate for adaptive loci was greater than 62% in all simulations. Figure S1B displays the map of a selection gradient obtained by rotation of 45 degrees from the reference axis, and Figure S1C displays the map of a selection gradient collinear to the direction of expansion.

~~Our approach modeled population structure at loci lacking or with weak association to environmental gradients in a way that reproduced the demographic model accurately.~~

In summary, adaptive loci were simulated so that the geographic distribution of the derived allele frequency correlated with the geographic distribution of the observed ecological gradient. Thus we modeled a situation in which allele frequencies at adaptive loci are truly associated with the observed ecological gradient and exhibit background population structure similar to allele frequencies at neutral loci.

## The orientation of ecological gradients impacts the identification of adaptive loci

To demonstrate universal properties of methods rather than showing differences in their relative performances, we focused on three methods that can be considered representative of ecological association tests. Our first method was based on a simple linear regression model, our second method used a linear model including correction for population structure based on the first principal component of the genotypic matrix, and our third method was based on latent factor mixed models.

Based on simulated data, we computed a false discovery rate (FDR) for the three association tests and for 17 distinct orientations of ecological gradients. For each simulation, the FDR was computed as the number of times that a positive test detected a selectively neutral locus. We used Bonferroni correction for multiple testing at a nominal type I error of 1%. For all ecological association methods, we found that all test performances were influenced by the angle between the selection gradient and the main axis of range expansion (Figure 2).

When we applied linear regression models and used Bonferroni correction, the FDR remained greater than 60% for all orientations of the ecological gradients (orange curve in Figure 2~~, Figure 3~~). For standard regression models, the excess of extreme $P$-values could be explained by the test not being correctly calibrated (see Figure S2). Including a correction for population structure did not improve the performances of linear regression methods. In contrast, increasing the cut-off threshold for multiple testing correction to $-\log_{10} P > 20$ or correcting for genomic inflation using the GIF improved the performances of

the regression models substantially (red curve in Figure 2). Using an increased cut-off threshold, the FDR curve exhibited a minimum at angular values close to zero, and the FDR was around 20% for those values (Figure 2, red curve). These results show that some gradient orientations are more favorable for detecting adaptive loci than other orientations. The most favorable case clearly occurs when selection gradients align with the main direction of range expansion. In more specific terms, this occurs when the orientation of the main axis of neutral genetic variation – as described by the first principal component of the neutral genotypic matrix – is perpendicular to the environmental predictor (François et al. 2010, Arenas et al. 2013, DeGiorgio and Rosenberg 2013). Ecological association methods perform well if they reduce the confounding effect created by population structure by capturing a significant proportion of genetic variation at neutral alleles in the residual error of the regression model.

Latent factor mixed models provide a general approach to correct for the undesired effect of population structure in ecological association methods (Frichot et al. 2013). In our simulations the number of hidden factors was chosen in order to make the distribution of $P$-values closest to a uniform distribution. We found that this value was around $K = 8$, consistent with the number of clusters detected by Bayesian clustering programs `STRUCTURE` and `TESS` (Pritchard et al. 2000, Chen et al. 2007). For this value of the number of factors, the performance of `LFMM` was expected to be superior to those of classic regression models whatever the orientation of the selection gradient. The FDR curve confirmed this expectation (Figure 2). Consistently with other regression models, the most favorable case happened when selection gradients aligned along the

direction of range expansion. Our explanation is that nominal error rates are better calibrated when selection gradients are parallel to the axis of expansion than in other orientations.

Extending our results to other methods, we observed that logistic regression models led to results very similar to those of linear regression models. While we reported results for models including the first principal component of the genotypic matrix, we observed that including more components led to qualitatively equivalent results. We did not use `BAYENV` because of high run-to-run variability. This inherent variability makes conclusions about genome-wide patterns of adaptation more difficult than for other methods (Blair et al. 2014). Since the `BAYENV` approach is similar to a logistic regression method including correction for population structure, the same behavior is expected for `BAYENV` as for the other methods investigated here.

To investigate the power of ecological association tests under various orientations of ecological gradients, we applied the Benjamini-Hochberg algorithm to control the FDR at level $q = 10\%$ for all methods. For each data set, we evaluated the sensitivity of tests as the proportion of loci with positive tests among adaptive loci. Sensitivity was generally high for linear regression methods, but the observed FDR reached values greater than 90% for those tests (e.g., Figure 3A and Figure 3B that shows excessively large numbers of neutral loci with $P$-values above the threshold). In contrast, the testing procedure based on latent factor models provided reasonable control of the FDR using the Benjamini-Hochberg algorithm. Note that the increased performance of `LFMM` resulted from the use of GIF corrections combined with the meta-analysis of multiple

runs, which is computationally more intensive than running simple regression models. For data sets containing 100 adaptive loci, the observed FDR ranged between 5% and 35% (10% expected, Figure 4A). For data sets containing 500 adaptive loci, the observed FDR ranged between 3% and 23% (Figure 4B). Using `LFMM`, the observed FDR was closer to its expected value when ecological gradients aligned along the direction of range expansion than along any other directions. The test power reached values greater than 75%, and it increased when ecological gradients aligned along the direction of range expansion. Power was less than 60% when ecological gradients were approximately perpendicular to the direction of range expansion (Figure 4).

# Discussion

Range expansions following climatic or other environmental changes are commonly associated with adaptive changes within migrant species genomes. This happens frequently in cases of species invasions (Kirk et al. 2013), postglacial recolonization (Hewitt 1999), and even in crop or animal domestication (Doebley et al. 2006). Researchers can investigate these changes by applying genome scan methods based on association with ecological gradients (Joost et al. 2007, Hancock et al. 2008, Jay et al. 2012). Here we provide new insights into the use of ecological association methods when species have expanded their spatial range. Our simulation study addressed the intuitive idea that the orientation of allele frequency gradients in geographic space could reveal signatures of natural selection (e.g., Fix 1996). If allele frequency gradients perpendicular to the axis of expansion can be linked to neutral population structure, allele frequency gra-

dients that align along the direction of expansion could be linked to selection. We found that ecological association approaches are useful ways to formalize these conceptual ideas.

Using spatially explicit simulations, our first result was that association tests are sensitive to the orientation of ecological gradients relative to the main axis of expansion. We found that the angle made by the axis of expansion and the axis of selection had a strong influence on the FDR (and power) of ecological association tests.Even for the best method, we observed up to 35% FDR in unfavorable cases. While we kept the origin of expansion fixed and modified the direction of selection gradients, our approach also applies to the symmetric case where a fixed ecological gradient is considered and the geographic origin of expansion is modified. Our second result is that the list of candidate genes obtained from association methods contained fewer false associations when the test variables exhibited geographic gradients that paralleled the direction of expansion than in other directions. Though the performance of methods could differ significantly, all association tests exhibited better performance when the selection gradient was parallel to the axis of expansion. When this gradient was orthogonal to the axis of expansion, the FDR increased in all methods.

The reason for the lower rates of false positives in the case of a North-South ecological gradient was that population genetic structure was organized West-East. Under population genetic models of range expansion and a broad set of conditions, the gradient of principal component maps are oriented along a direction perpendicular to the axis of the expansion, rather than parallel to expansion (François et al. 2010, De Giorgio and Rosenberg 2013). This pattern

is an outcome of the "allele surfing" phenomenon, which creates patches of high allele-frequency differentiation that align perpendicular to the direction of the expansion, and complicates the detection of selection when ecological gradients do not align with the expansion axis. Francois et al (2010) suggest that the results presented here will be valid for geometries more complex than a rectangular array, for example, range expansions in the European continent. They also suggest that admixture events have an impact on test performances. De Giorgio and Rosenberg (2013) provided evidence that population sampling can modify principal component analysis, which impacts the power and FDR of tests. While the allele surfing phenomenon may strongly bias allele-frequency differentiation tests, we observed that the undesired effects can be corrected when we sample throughout the whole species range and test gradients that are perpendicular to the first principal axis of neutral variation.

The most likely explanation for the high FDRs observed in linear regression models is that those methods use an incorrect model to test the null hypothesis. In the linear model, residual errors are considered statistically independent of each other ignoring population genetic structure shaped by range expansion (Figure 3, Figure S2). Using corrections based on principal components improved the FDR only slightly, and the results were qualitatively similar to those of linear regression methods. We also observed that increasing the number of principal components reduced the power to reject the null hypothesis (not reported).

The FDR of linear regression methods decreased substantially when an ultra-conservative cut-off threshold defined the test significance level. Again, we found

that conservative tests exhibited better performances when ecological gradients paralleled the expansion axis. Because the calibration of $P$-values based on linear models is usually incorrect, researchers must be cautious about interpretations of the candidate locus list. In addition they must be aware that their results come without any control of the FDR.

Latent factor mixed linear models were associated with much lower levels of FDR than simple linear regression models and models using correction based on principal components. Instead of principal components, `LFMM` uses unknown factors in addition to the fixed environmental effects. The `LFMM` algorithm estimates the unknown factors from the data at the same time as it estimates the effect of ecological variables. Choosing the number of factors according to the flatness of the $P$-values histogram, as measured by the genomic inflation factor, provides assurance that association tests were correctly calibrated, and models of background variation (or correlated residual errors) remained at acceptable levels. An important advantage of this choice procedure is to allow researchers to analyze ranked lists of loci while they control the FDR using classical procedures (Benjamini & Hochberg 1995). Greater power will be found in ecological gradients that parallel the axis of expansion, and fewer false discoveries will be done in data sets where the number of adaptive loci is high compared to the number of neutral loci.

Recent simulation studies have evaluated the sensitivity and specificity of genome scans for selection, and compared ecological association methods to methods based on allele frequency differentiation (De Mita et al. 2013, de Villemereuil et al. 2014). These simulation studies have shown that ecological

methods have higher power to detect adaptive genetic variation than outlier-based methods when adaptive traits are influenced by several genes and when population structure is hierarchical (de Villemereuil et al. 2014). Our study confirmed that ecological association methods could detect adaptive genetic variation when populations have undergone range expansion, and have high power to detect adaptive genetic variation when expansion and ecological gradients follow the same direction.

Following our observations, we encourage researchers to use ecological association methods when screening genomes for local adaptation in spatially expanding populations. For example, ecological gradients that align along the axis of expansion occur for species that colonized Europe following the most recent glacial period (Hewitt 2000). More generally, our results are particularly relevant to global change scenarios where species track their ecological niche during range expansion while adapting to changing environments at their rear edge (Jump & Penuelas 2005). In this case, researchers using ecological association approaches should be aware that detecting genomic signature of adaptation can be facilitated when gradients align along the main axis of expansion.

# Acknowledgments

EF was supported by a grant from "la Région Rhône-Alpes" (CIBLE2011). SDS was supported by the National Science Foundation (OISE-0965038). OEG was supported by the Marine Alliance for Science and Technology for Scotland (MASTS). OF acknowledges support from Grenoble INP and Persyval-lab.

# References


Arenas M, François O, Currat M, Ray N, Excoffier L. (2013). Influence of admixture and Paleolithic range contractions on current European diversity gradients. *Molecular Biology and Evolution* **30**:57-61.

Blair LM, Granka JM, Feldman MW. (2014). On the stability of the Bayenv method in assessing human SNP-environment associations. *Human Genomics* **8**:1.

Blois JL, Hadly EA. (2009). Mammalian response to Cenozoic climatic change. *Annual Review of Earth and Planetary Sciences*, **37**:181-208.

Brown, M. (1975). A method for combining non-independent, one-sided tests of significance. *Biometrics* **31**: 987-992.

Buckley J, Butlin RK, Bridle JR. (2012). Evidence for evolutionary change associated with the recent range expansion of the British butterfly, *Aricia agestis*, in response to climate change. *Molecular Ecology* **21**: 267-280.

Benjamini Y, Hochberg Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**: 289-300.

Chen C, Durand E, Forbes F, François O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes* **7**: 747-756.

Comes HP, Kadereit JW. (1998). The effect of Quaternary climatic changes on plant distribution and evolution. *Trends in Plant Sciences* **3**: 432-438.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185**:

1411-1423.

Currat M, Ray N, Excoffier L. (2004). SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes* **4**:139-142.

Davis MB, Shaw RG. (2001). Range shifts and adaptive responses to quaternary climate change. *Science* **292**: 673-679.

Davis MB, Shaw RG, Etterson JR. (2005). Evolutionary responses to changing climate. *Ecology* **86**: 1704-1714.

De Giorgio M, Rosenberg NA. (2013). Geographic sampling scheme as a determinant of the major axis of genetic variation in principal components analysis. *Molecular Biology and Evolution* **30**: 480-488.

De Mita S, Thuillet A-C, Gay L, Ahmadi N, Manel S, Ronfort J, Vigouroux Y. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology* **22**: 1383-1399.

De Villemereuil P, Frichot E, Bazin E, Francois O, Gaggiotti OE. (2014). Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology* **23**: 2006-2019.

Devlin B, Roeder K. (1999). Genomic control for association studies. *Biometrics* **55**: 997-1004.

Doebley JF, Gaut BS, Smith BD. (2006). The molecular genetics of crop domestication. *Cell* **127**: 1309-1321.

Eckert AJ, Bower AD, Gonzalez-Martinez SC, Wegrzyn JL, Coop G, Neale DB. (2010). Back to nature: Ecological genomics of Loblolly pine (*Pinus taeda,*

*Pinaceae*). *Molecular Ecology* **19**: 3789-3805.

Edmonds CA, Lillie AS, Cavalli-Sforza LL. (2004). Mutations arising in the wave front of an expanding population. *Proc Natl Acad Sci USA* **101**:975-979.

Felsenstein J. (2002). Contrasts for a within-species comparative method. In "Modern Developments in Theoretical Population Genetics", ed. M. Slatkin and M. Veuille, pp. 118-129, Oxford University Press, Oxford.

Fix AG. (1996). Gene frequency clines in Europe: Demic diffusion or natural selection? *The Journal of the Royal Anthropological Institute* **4**: 625-643.

François O, Currat M, Ray N, Han E, Excoffier L, Novembre J. (2010). Principal component analysis under population genetic models of range expansion and admixture. *Molecular Biology and Evolution* **27**: 1257-1268.

Frichot E, Schoville SD, Bouchard G, François O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution* **30**: 1687-1699.

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics* **7**:e1002355.

Grafen A. (1989). The phylogenetic regression. *Philosophical Transaction of the Royal Society of London Series B, Biological Sciences* **326**:119-157.

Haldane JBS. (1948). The theory of a cline. *Journal of Genetics* **48**:277-284.

Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK et al. (2008). Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics 4*: e32.

Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, et

al. (2010). Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci USA 107*:8924-30.

Harvey PH, Pagel MD. *The Comparative Method in Evolutionary Biology*. Oxford, UK: Oxford University Press, 1991.

Hermisson J, Pennings PS. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**: 2335-52.

Hermisson J. (2009). Who believes in whole-genome scans for selection. *Heredity* **103**: 283-284.

Hewitt GM. (1999). Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society* **68**: 87-112.

Hewitt GM. (2000). The genetic legacy of the quaternary ice ages. *Nature* **405**: 907-913.

Jay F, Manel S, Alvarez N, Durand EY, Thuiller W, Holderegger R, Taberlet P, François O. (2012). Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology* **21**: 2354-2368.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, et al. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55-61

Joost S, Bonin A, Bruford MW, Després L, Conord C, Erhardt G, Taberlet P. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology* **16**: 3955-3969.

Jump AS, Penuelas J. (2005). Running to stand still: adaptation and the

response of plants to rapid climate change. *Ecology Letters* **8**: 1010-1020.

Kirk H, Dorn S, Mazzi D. (2013). Molecular genetics and genomics generate new insights into invertebrate pest invasions. *Evolutionary Applications* **6**:842-856.

Parmesan C, Yohe G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. *Nature* **421**: 37-42.

Poncet BN, Herrmann D, Gugerli F, Taberlet P, Holderegger R, Gielly L, Rioux D, Thuiller W, Aubert S, Manel S. (2010). Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Molecular Ecology* **19**: 2896-2907.

Pritchard JK, Stephens M, Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945-959.

Pritchard JK, Pickrell JK, Coop G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology* **20**:R208-15.

Schoville SD, Bonin A, François O, Lobreaux S, Melodelima C, Manel S. (2012). Adaptive genetic variation on the landscape: methods and cases. *Annual Review of Ecology, Evolution and Systematics* **43**:23-43.

Teshima K, Coop G, Przeworski MF. (2006). How reliable are empirical genomic scans for selective sweeps? *Genome Research* **16**: 702-712.

Vitti JJ, Grossman SR, Sabeti PC. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics* **47**: 97-120.

Yoder JB, Stanton-Geddes J, Zhou P, Briskine R, Young ND, Tiffin P. (2014). Genomic signature of local adaptation to climate in *Medicago trun-*

*catula*. *Genetics* **196**: 1263-1275.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF,McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**:203-208.

White TA, Perkins SE, Heckel G, Searle JB. (2013). Adaptive evolution during an ongoing range expansion: the invasive bank vole (*Myodes glareolus*) in Ireland. *Molecular Ecology* **22**: 2971-2985.

# Figure legends

**Figure 1.** Schematic representation of evolutionary scenarios. Populations (demes) are represented by a regular array of dots, the larger ones indicating the origin of expansion. The main direction of expansion is shown by black arrow (solid line), the circular wave front is shown by an orange circle. The main axis of the ecological gradient is shown by a green arrow (dashed line) which angle varies from -90 degrees to +90 degrees.

**Figure 2.** False discovery rate (FDR) for ecological association tests. Values of the FDR as a function of the orientation of ecological gradients relative to the main axis of range expansion. We used 17 angle values varying from -90 degrees to +90 degrees and three association tests: simple linear regression, principal component regression using the first PC of the genotypic matrix and a latent factor model using $K = 8$ latent factors.

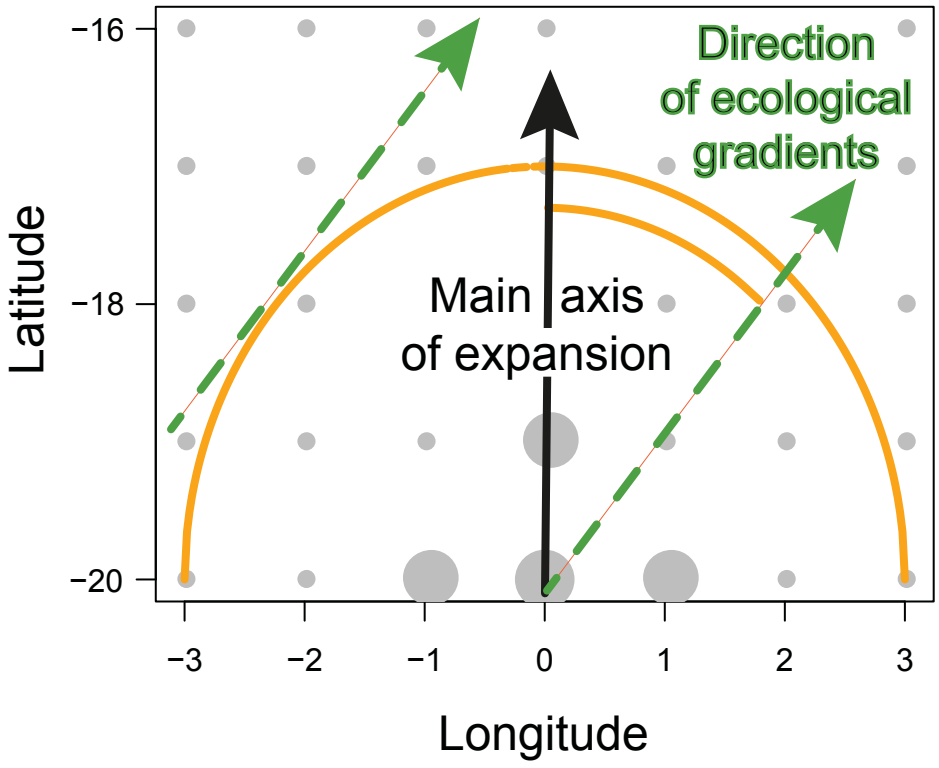
**Figure 3.** Manhattan plot for linear model association tests under two distinct angles of selection gradient (0 and 90 degrees). Graphical representation of minus $\log_{10} P$-values of linear regression tests at each locus. The horizontal line represents the value of the Bonferroni correction threshold. $P$-values for A) ecological gradients perpendicular to the direction of expansion and B) parallel to the direction of expansion.

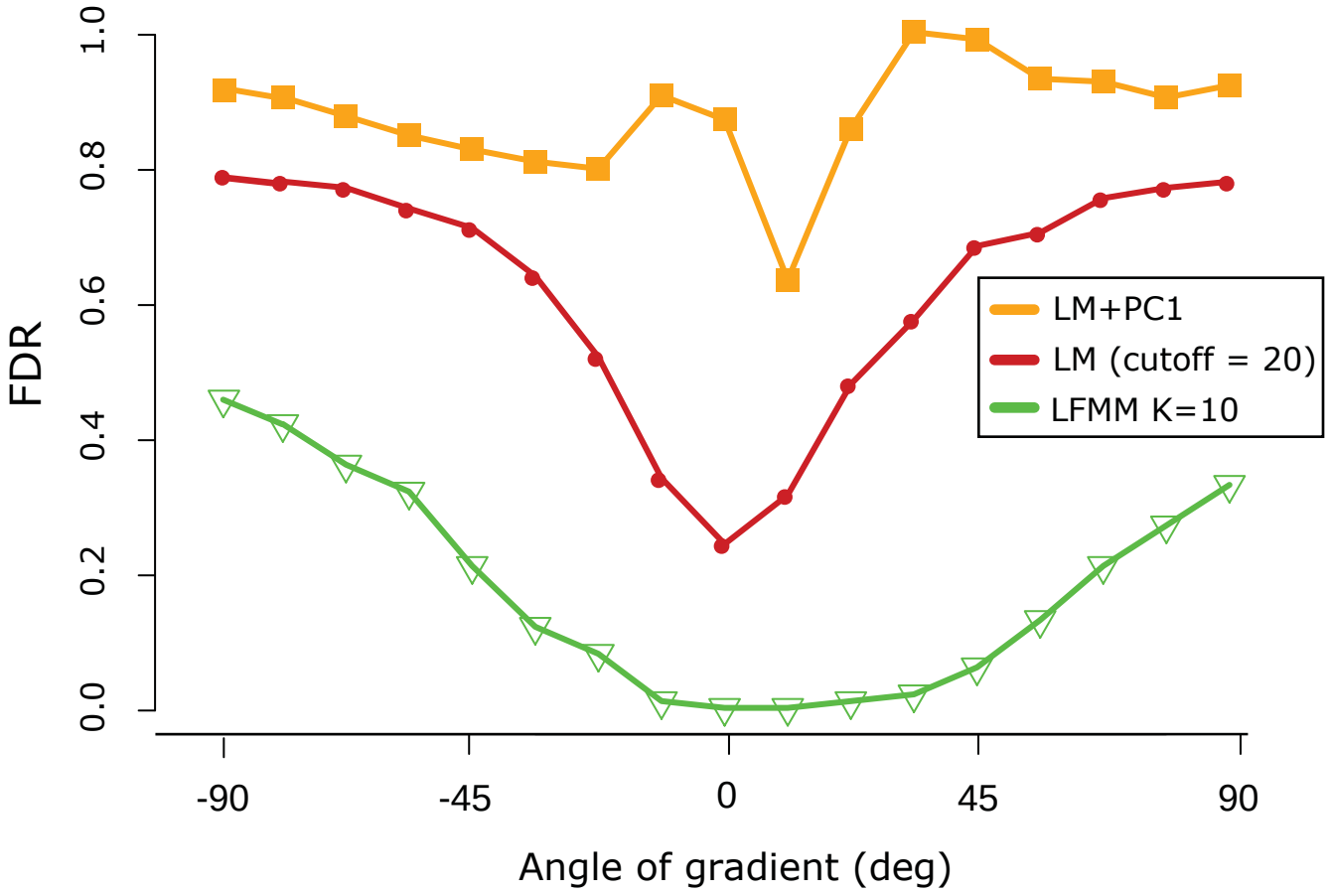
**Figure 4.** False discovery rate (FDR) – Power plot for `LFMM` tests and 17

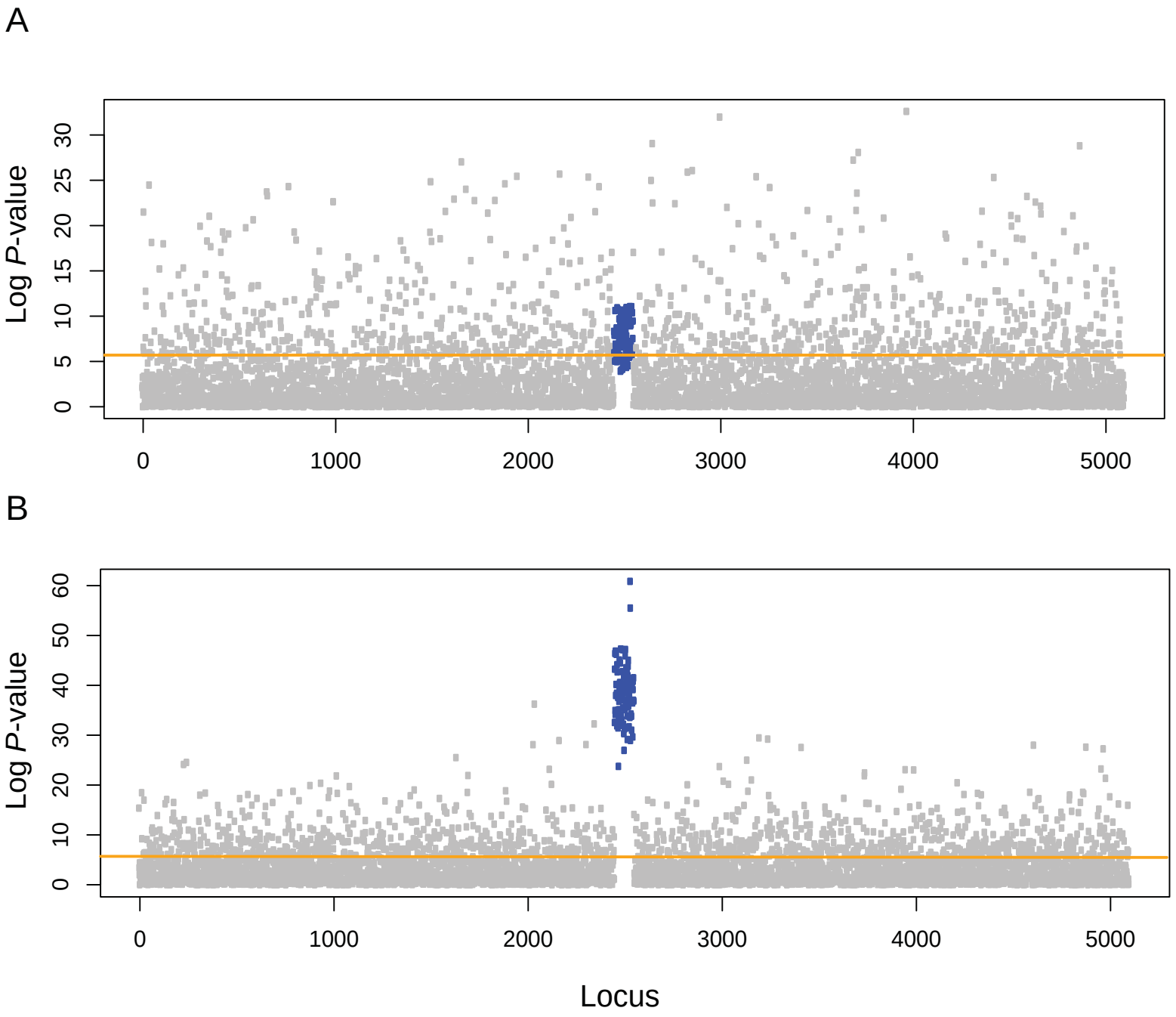
distinct orientations of ecological gradients. Each data set is represented by an arrow displaying the direction of the ecological gradient in the simulated data. Vertical arrows indicate that the ecological gradient aligns along the direction of expansion. The expected FDR, $q = 10\%$, is shown by a vertical line. Each arrow position corresponds to the sensitivity (power) of tests and the percentage of false discoveries in the lists of loci obtained with the Benjamini-Hochberg algorithm. A) 100 adaptive loci, B) 500 adaptive loci. Five runs and $K = 8$ factors were used in `LFMM`.

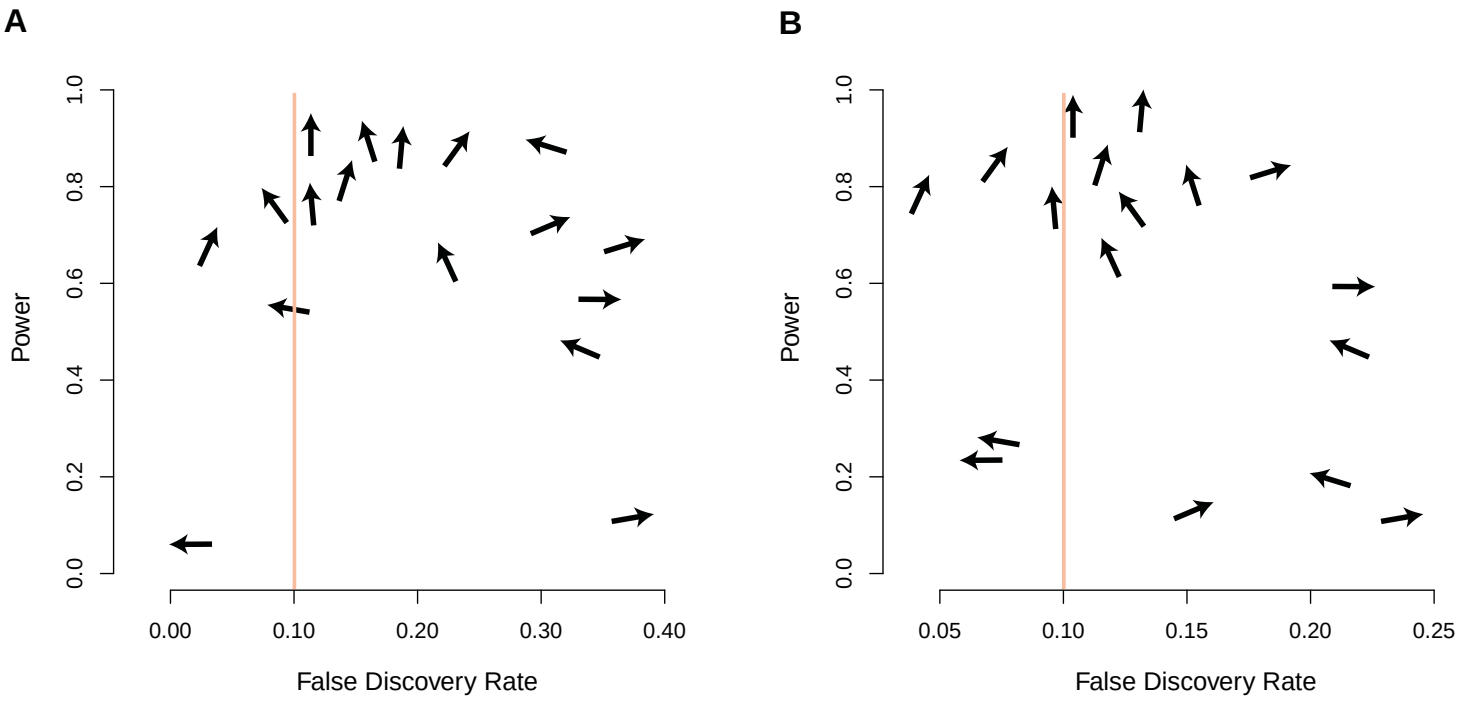
**Figure S1.** Neutral genetic variation and ecological gradients in simulations scenarios. A) Principal component analysis: Map of the first principal component of neutral genetic variation. The axis of expansion is indicated by a vertical arrow (solid line). B) Map of selection gradient. The direction of selection gradient forms an angle of 45 degrees with the direction of expansion (dashed line). C) Second example of a map of selection gradient. The selection gradient aligns to the direction of expansion (angle of 0 degree), and it is perpendicular to the first axis of genetic variation.

**Figure S2.** Empirical distribution of $P$-values for the linear model and for latent factor mixed models. Data simulated for ecological gradients that align to the direction of expansion ($K = 8$ factors were used in `LFMM`).


Direction
of ecological
gradients
Main axis
of expansion
Latitude
−16
−18
−20
−3
−2
−1
0
1
2
3
Longitude


FDR
0.0
0.2
0.4
0.6
0.8
1.0
-90
-45
0
45
90
Angle of gradient (deg)
LM+PC1
LM (cutoff = 20)
LFMM K=10


A
Log P-value
0
5
10
15
20
25
30
0
1000
2000
3000
4000
5000
B
Log P-value
0
10
20
30
40
50
60
0
1000
2000
3000
4000
5000
Locus


A
Power
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.10
0.20
0.30
0.40
False Discovery Rate
B
Power
0.0
0.2
0.4
0.6
0.8
1.0
0.05
0.10
0.15
0.20
0.25
False Discovery Rate